표본자료분석

– 가중치, 복합표본조사자료분석

4. 가중치

- **가중치를 부여하는 이유**
  - **불균등 선택 확률(unequal selection probability)을 보정하기 위해**
  - **무응답(non-response)을 보정하기 위해**
  - 모집단에서 이미 알려져 있는 특정 변수 의 분포(예; 성, 연령, 지역 등)와 표본 결과를 일치시키기 위해 조정하는 **벤치마킹(bench-marking) /사후층화(post-stratification)를** 위해
    - 표본 추정치의 정도를 향상시키기 위해
    - 무응답 및 비포함(non-coverage) 표본틀을 보정하기 위해

- **가중치 반영 추정의 장점**
  - 영역별로 서로 다른 추출률로 표본을 추출하여 모집단 크기나 표본추출률이 작거나 낮은 영역의 추출률을 보정하므로 전체 정도가 향상
    - 편향 제거
    - 분산 증가
  - 중복, 비포괄성 등의 불완전한 추출틀 문제 해결
  - 조사에서 발생되는 무응답 문제를 해결

- **최종 가중치의 표현과 부여 과정**
  - **불균등 선택 확률의 가중치** w_1
    - 표본추출단위와 관련된 실제 표본의 선택 과정을 반영
  - **표본의 무응답 조정을 위한 가중치** $w_{2.1}$
    - 무응답률 반영
    - 무응답으로 인한 원표본(original sample)과의 불균형을 조정하기 위해 가중 층 조정(weighting class adjustment :WCA)치를 계산하여 반영
  - **모집단의 비포함(non-coverage) 및 분포 조정을 위한 사후층화 가중치** $w_{3.21}$
    - 중요한 의미를 지닌 특정 변수에 대한 모집단 분포와 표본 분포의 불균형을 수정하기 위해 사후층화 조정(PCA) 가중치를 계산하여 반영
  - **전체 가중값** $w = w_1 \times w_{2.1} \times w_{3.21}$

- 기본 가중치 산출 과정

▪ 가중치 = 추출율 역수 × 응답률 역수

1. 추출율 역수 = $\frac{\text{모집단조사구수}}{\text{표본조사구수}} \times \frac{\text{적절가구수}}{\text{표본가구수}}$

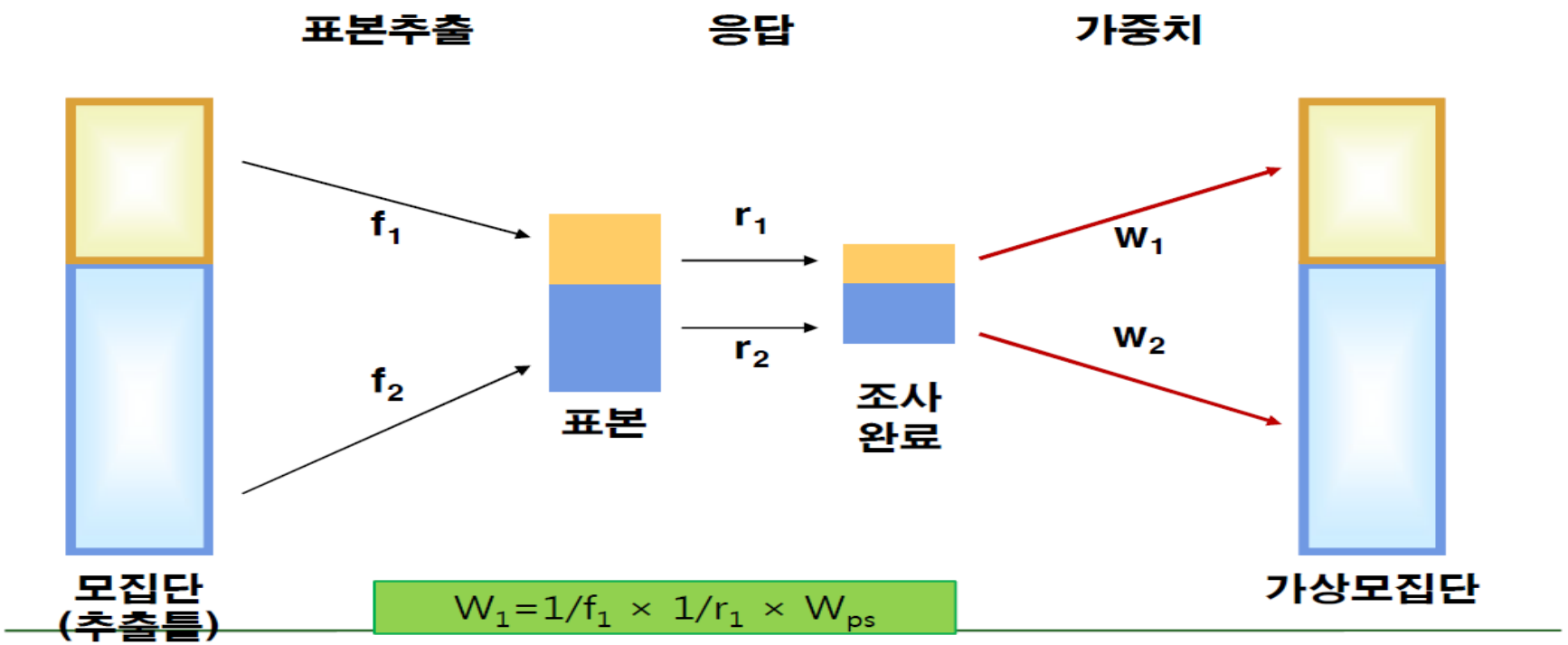
2. 응답률 역수 = $\frac{\text{조사대상가구수}}{\text{참여가구수}} \times \frac{\text{조사대상가구원수}}{\text{참여가구원수}}$

3. 가중치 사후보정

   - 모집단의 인구구성비와 맞춤

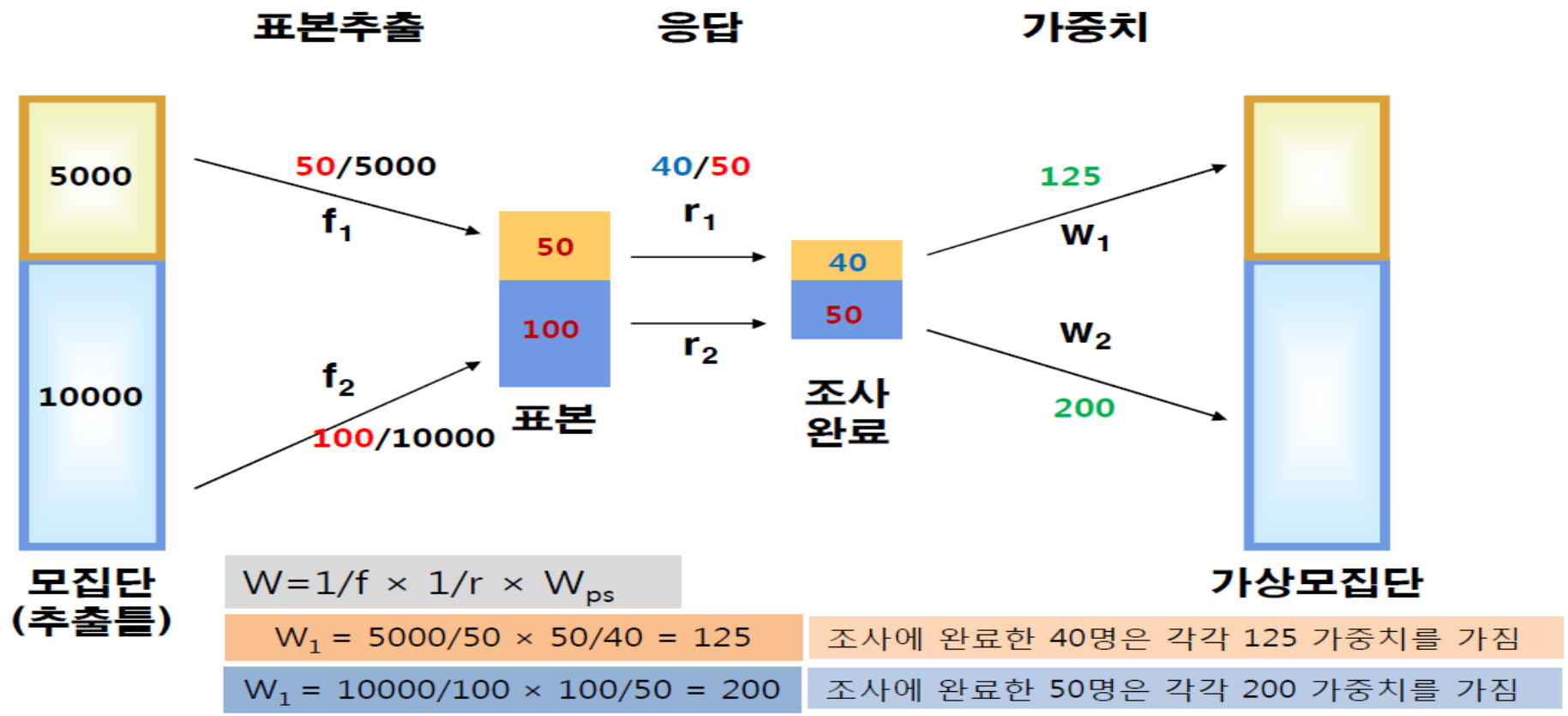
$$\text{최종 가중치} = \text{가중치} \times \frac{\text{인구(성별 · 연령별)}}{\text{가중치합 (성별 · 연령별)}}$$

▪ 조사에 참여한 표본이 모집단을 대표하도록 가중치를 부여



인용 : 국민건강영양조사

- **예 : 가중치 산출**



$W = 1/f \times 1/r \times W_{ps}$

$W_1 = 5000/50 \times 50/40 = 125$ 조사에 완료한 40명은 각각 125 가중치를 가짐

$W_1 = 10000/100 \times 100/50 = 200$ 조사에 완료한 50명은 각각 200 가중치를 가짐

인용 : 국민건강영양조사

예) 기본 가중치 산출 결과(1) : 가구 및 가구원 가중치

표본가구 추출율 계산

조사구	조사구 (지역,층,주거유형)	모집단 조사구	표본 조사구	적절 가구수	표본 가구수	조사구 추출	가구 추출	최종 추출
A115	서울, 1, 일반	30,927	19	67	23	1627.7	2.9	4741.67

$$\text{추출율 역수} = \frac{30,927}{19} \times \frac{67}{23} = 1627.7 \times 2.9 = 4741.67$$

개인 응답률 계산

조사구	개인	대상 가구수	참여 가구수	대상 가구원	참여 가구원	가구 응답	가구원 응답	최종 응답
A115	010101	23	21	3	2	1.1	1.5	1.64

$$\text{응답률 역수} = \frac{23}{21} \times \frac{3}{2} = 1.1 \times 1.5 = 1.64$$

가중치 계산

A115-010101 가중치 = 4741.67 x 1.64 = 7789.88

- **기본 가중치 산출 결과(2) : 성별/연령대별 가중치**

성별·연령별 가중치 합

연령	남	여
01-04	1,243,576	989,751
05-09	1,968,326	1,740,850
10-14	2,006,804	1,755,239
…	…	…
70-74	989,938	1,335,956
75-	716,239	1,384,914

모집단의 성별·연령별 인구

연령	남	여
01-04	963,710	891,935
05-09	1,563,013	1,427,565
10-14	1,822,845	1,615,614
…	…	…
70-74	581,198	799,808
75-	544,888	1,100,386

$$\text{최종가중치} = \text{가중치} \times \frac{\text{인구 (성별·연령별)}}{\text{가중치합 (성별·연령별)}}$$

$$= 7{,}789.88 \times \frac{963{,}710}{1{,}243{,}576}$$

$$= 6036.78$$

- 예 : 병원을 추출하는 경우
  - 추출방안

    1) SRS : 병원규모를 무시하고 동일한 추출확률로 추출하는 방안

    - N=10개이므로 1~10까지의 난수 중 3개(1, 2, 10)를 확률적으로 추출하여 표본 추출

    2) SYS_PPS : 병원 의사수(혹은 환자수)를 반영해 병원마다 다른 추출확률로 계통추출하는 방안(의사수 기준)

    - 총 의사수=328명이므로 1~109까지의 난수 86이 추출되면 195, 304가 속한 병원을 표본으로 추출

id	의사수	누적합	매출액	추출률	표본	설계 가중치		무응답조정후 가중치			사후층화조정 후 최종 가중치	
						PPS	SRS	응답	PPS	SRS	PPS	SRS
10	91	91	72411	0.2774	O	1.2015	3.3333	O	1.8022	5	1.25	5
1	128	219	53100	0.3902	O	0.8542	3.3333	X				
8	39	258	23824	0.1189								
9	22	280	5800	0.0671								
6	6	286	4064	0.0183								
3	6	292	2797	0.0183								
2	13	305	2757	0.0396	O	8.4103	3.3333	O	12.6164	5	8.75	5
4	4	309	2200	0.0122								
5	8	317	1950	0.0244								
7	11	328	1849	0.0335								
계	328	328	170752.1	1	3	10.4659	10		14.4176	10	10	10

5. 복합표본조사자료 분석 패키지 소개

- **복합표본조사자료의 특징**
  - 층화추출, 집락추출, 다단계추출의 과정으로 추출된 표본을 대상으로 수행된 표본조사를 의미
    - 추출 확률이 서로 다른 표본이 존재
  - 단순확률추출법의 분석 방법의 적용은 편향된 결과를 제공하므로 표본추출과정 및 가중치를 이용한 표본분석이 필요
    - 모수 추정 과정에서 추출법과 가중치의 반영 여부는 모수 추정 결과의 편향여부에 영향 미침
    - 추정 결과를 이용한 검정 결과의 편향 원인 제공
  - 복합표본조사자료를 분석하는 패키지 사용이 필요
    - 일반통계패키지 사용 시 모수 추정은 비편향 결과를 얻을 수 있으나 분산 추정은 편향된 결과를 얻게 됨에 주의

예 : 모총합 추정 과정

- **PPS :** $\sum_{i=1}^{3} w_i x_i = 1.1480 \times 91 + 0.8161 \times 128 + 8.0359 \times 13 = 313$
- **SRS :** $\sum_{i=1}^{3} w_i x_i = 3.3333 \times 91 + 3.3333 \times 128 + 3.3333 \times 13 = 773$

id	의사 수	매출액	표본	최종 가중치		매출액(추정)		의사 수(추정)	
				pps	SRS	pps	SRS	pps	SRS
10	91	72411	O	1.1480	3.3333	83127	241370	104.5	303.3
1	128	53100	O	0.8161	3.3333	43337	177000	104.5	427.7
8	39	23824							
9	22	5800							
6	6	4064							
3	6	2797							
2	13	2757	O	8.0359	3.3333	22155	9190	104.5	43.3
4	4	2200							
5	8	1950							
7	11	1849							
계	328	170752		10	10	148619	427560	313.4	773.3

복합표본조사분석(1) : SAS

- **survey프로시져**

추정	단순임의추출방법 (Simple Random Sampling Design)	복합표본설계추출방법 (Complex Sampling Design)
평균	Proc means data=a; Var HE_BMI; Run;	Proc surveymeans data=a; Strata Kstrata; Cluster PSU; Weight WT_ex; Var HE_BMI; Run;
비율	Proc freq data=a; Table HE_OBE; Run;	Proc surveyfreq data=a; Strata Kstrata; Cluster PSU; Weight WT_ex; Table HE_OBE; Run;

인용 : 국민건강영양조사

회귀 분석	단순임의추출방법 (Simple Random Sampling Design)	복합표본설계추출방법 (Complex Sampling Design)
연속	Proc reg data=a; Model HE_BMI =sex Run;	Proc surveyreg data=a; Strata Kstrata; Cluster PSU; Weight WT_ex; Model HE_BMI=sex; Run;
명목 (로지스틱)	Proc logistics data=a; Model HE_BMI25(event='1')=sex; Run;	Proc surveylogistic data=a; Strata Kstrata; Cluster PSU; Weight WT_ex; Model HE_BMI25(event='1')=sex; Run;

인용 : 국민건강영양조사

- **SAS 분석 결과**

프로그램	결과
Proc means data=HN10_all N MEAN STDERR; Var HE_BMI; Run;	N: 8407, 평균값: 22.2910466, 표준오차: 0.0442337
Proc means data=HN10_all N MEAN STDERR; Weight wt_itvex; Var HE_BMI; Run;	N: 8407, 평균값: 22.6143336, 표준오차: 0.0432909
Proc surveymeans data=HN10_all NOBS MEAN STDERR; Strata kstrata; Cluster PSU; Weight wt_itvex; Var HE_BMI; Run;	N: 8407, 평균값: 22.614334, 표준오차: 0.058838

인용 : 국민건강영양조사

복합표본조사분석(2) : SPSS

- 복합표본(complex samples)

▪ 분석 -> 복합표본

→ 1. 분석 준비(복합표본 계획파일 생

→ 2. 평균, 유병률, 카이제곱분석

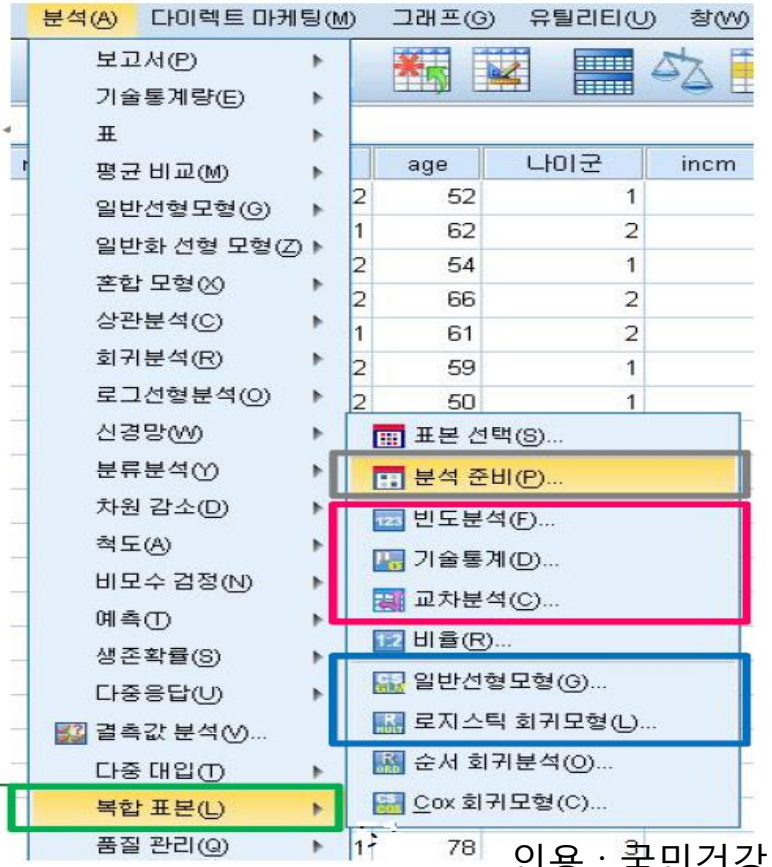
-기술통계(평균)

-교차분석(카이제곱분석)

2. 연관성 분석

- 일반선형모형 (회귀분석)

- 로지스틱회귀분석

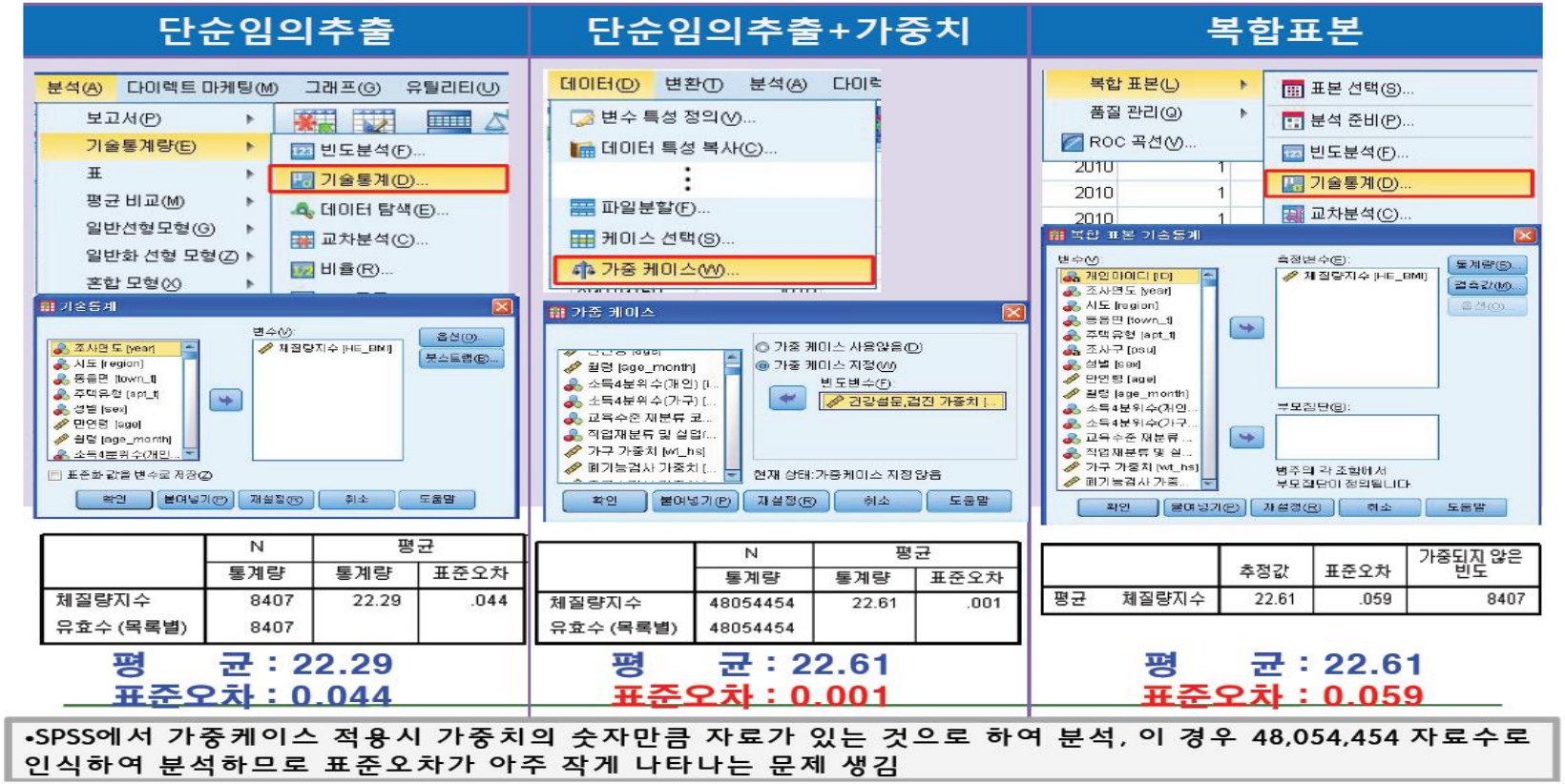


• 분석 -> 복합표본 -> 분석준비에서 복합표본설계요소(계층, 군집, 표본가중값) 지정하고 그 파일을 CSPLAN파일로 저장

• 분석은 분석 -> 복합표본 -> 로지스틱회귀모형(예, e.g.) 과 생성된 CSPLAN파일를 사용하여 분석

인용 : 국민건강영양조사

- **SPSS 복합표본 분석 결과**



인용 : 국민건강영양조사

- **R survey package** :
  - R project(R foundation)에서 개발
  - 분석 가능한 표본 : 층화추출, 집락추출, 다단계추출,
    불균등 추출 확률 및 가중치를 갖는 표본설계
    - ✓R pps package, survey function
  - Descriptive, GLM, 생존분석(비례위험모형)
  - 무응답보정, 사후층화추정과 Raking 가중치 계산 가능
  - 분산추정 : 선형화, 반복 가중치를 이용한 추정